Proposed new course: Technical Tools for Linguists
Course Number: Linguistics 5050

Instructor Name: _________________________
Office: _________________________
Phone: _________________________

Meeting Date/Time:
Classroom Location:

Course description:
Practical training in standard computational tools for tackling different kinds of linguistic research. Students will learn computational techniques to access, search and format linguistic datasets, including text corpora, speech and audio, structured representations including parse trees, and experimental measurements. The course will also cover data exploration and basic modeling.

No prerequisite in programming is required: the course will cover the necessary skills. The course is designed to be hands-on, and students will have the opportunity to work on the problem sets during the class sessions.

Course goals, learning objectives/outcomes:

1. Students will gain hands-on experience gathering, formatting, and manipulating data.
2. Students will learn to use corpus, field, and experimental data, as well as to combine data from multiple sources.
3. Students will learn to work with existing computational tools.
4. At the end of the course, students will be able to process massive amounts of linguistic data.

The course is designed to stand alone, but also to provide an introduction to the graduate Computational Linguistics sequence. It is not a prerequisite for the Computational Linguistics courses, but is helpful for students who lack any prior experience with computational tools.

Content topic list

- Accessing and navigating corpora
- Linguistic data manipulation and visualization
- Automatic processing of structured linguistic representations
- R scripting
- Praat scripting

Required texts and Course materials

Books:
R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics*

using R. Cambridge University Press. [http://searchworks.stanford.edu/view/7520794]

Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. Chapter 11 [http://nltk.org/book/ch11.html]

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA.

Mark Pilgrim. 2000. *Dive Into Python*. [http://www.diveintopython.net/]

Will Styler. 2013. *Using Praat for linguistic research*. [http://savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf]

Articles:
Marie-Catherine de Marneffe, Scott Grimm, Inbal Arnon, Susannah Kirby, and Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1): 25-61.

Jiahong Yuan, Mark Liberman and Christopher Cieri. 2006. Language and gender differences in speech overlaps in conversation. *Journal of the Acoustical Society of America* 120.

Tutorials:
- R tutorial [http://www.cyclismo.org/tutorial/R/]
- Python tutorial [http://docs.python.org/3.0/tutorial/]
- Unix for poets [http://ufal.mff.cuni.cz/~hladka/tutorial/UnixforPoets.pdf]

Optional readings:
Peter Dalgaard. 2008. *Introductory Statistics with R, 2nd edition*. Springer. [http://searchworks.stanford.edu/view/7720505]

Keith Johnson. 2008. *Quantitative Methods in Linguistics*. Blackwell.

Assignments
The assignments are designed to be relevant to linguistic research, and will be connected to the work and interest of the students. There will be 5 assignments, one for each unit in the class. The assignments are described in detail above; each one will require students to write a short program to perform some analysis of a dataset (for instance, assignment 1 is to write a Python program measuring utterance lengths by men and women in a section of the Fisher corpus). Students will work on the assignments both in class and at home, and will be encouraged to work collaboratively in small groups, but everyone has to turn in his/her own assignment.

Syllabus

[weeks 1-3] Unit 1: Basic data manipulations

- Introduction and motivation
  - Case study: **Do women talk more than men?**
  - How to use dialogue corpora to test a hypothesis
- Basic Unix environment
  - How to access and navigate our corpora directories
- A computer language to deal with human language: Introductory Python
  - Basic file IO
  - Decision-making: logic, comparatives, conditionals

Readings:

- Jiahong Yuan, Mark Liberman and Christopher Cieri. 2006. Language and gender differences in speech overlaps in conversation. *Journal of the Acoustical Society of America* 120.
- Unix for poets
- *Dive Into Python*: chapter 2, chapter 6.2

Week 1 will discuss the basic goals and assumptions of data-driven corpus linguistics, giving a broad overview of what language corpora are available, how they are created, where to find them and what problems they might be helpful in solving. We will introduce the Fisher dialogue corpus and instruct students in the tools necessary to find, view and search files from Fisher by hand using the standard Unix environment. Students will do several in-class exercises designed to build competence with the Unix command line tool set, and will conduct a small corpus investigation of whether women or men talk more by hand-analyzing Fisher data.

Weeks 2-3 will introduce Python as a tool for programmatically reading and searching large amounts of data. We will explain Python's basic representations of numbers and strings. Students will gain experience using Python as a calculator and searching for words or phrases from the interactive Python shell. Students will learn the basic constructs of imperative programming (loops and conditionals).

Assignment 1 Students will have to write a Python program to decide the women-vs-men problem by counting utterance lengths in a gender-annotated dialogue corpus. Students will need to provide a set of statistics, coming out of their Python program, to answer the question of whether women talk more than men.

[weeks 4-6] Unit 2: Reading text and counting words

- Case study: **Investigating Zipf's law**
- Counting instances of a word in a file
- Dictionaries
- Counting all words/bigrams

Readings:

- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA – Chapter 1.
- *Dive Into Python*: chapter 3

Weeks 4-6 will focus on structured data types (lists and dictionaries). Students will learn to store and manage data, to develop appropriate data models for simple structured problems, and to write simple procedures for processing stored data.

Assignment 2 will be to write Python programs for counting unigrams and bigrams that can verify the statistical pattern of word frequencies known as Zipf's Law.

[weeks 7-10] Unit 3: Dealing with linguistic structured representations

- Case study: **Which verbs appear most often in the passive construction?**
- Field-structured: CSV, space-delimited
- Tree-structured: Penn TreeBank parses
- Structures with internal references: CoNLL parses
- XML parsing
- Internationalization, non-standard character sets:
  How to deal with Arabic, Chinese, Hindi or Cyrillic?

Readings:

- Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit.* O'Reilly Media. Chapter 11 [http://nltk.org/book/ch11.html]

Weeks 7-10 will extend the ability to use data structures to familiarize students with the use of external libraries for reading and manipulating common data formats. We will introduce the NLTK libraries for reading parse trees. Students will gain expertise at searching for and understanding documentation.

Assignment 3 will be to build Python word counting programs for several formatted datasets, using pre-developed libraries to read the data into appropriate representations and computing lists of verbs which tend to appear in passive constructions. If time allows, students will compare the suitableness of constituency and dependency trees as representations for such a study.

[weeks 11-12] Unit 4: R scripting

- Case study: **Dative alternation: how do children differ from adults?**
- The R language:
  - Variables, control statements and data structures in R
  - Data exploration and visualization

Readings:

- R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press – Chapters 1 and 2.

- Marie-Catherine de Marneffe, Scott Grimm, Inbal Arnon, Susannah Kirby, and Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1): 25-61.

Students will learn the basics of data storage and manipulation in R, applying their understanding of Python programming to "translate" basic concepts into a new language.

Assignment 4 will be to write a statistical analysis of the provided dataset using the language and create a visualization of the results.

[weeks 13-14] Unit 5: Praat scripting

- Case study: **Automatically extracting measurements to make vowel space plots**
- The Praat language
    - Variables, control statements and data structures in Praat
    - Manipulation of audio data

Readings:

- *Using Praat for linguistic research*: chapters 2, 3 and 11.

During the last two weeks, students will learn the basics of Praat scripting, applying their understanding of Python programming to "translate" basic concepts into a new language.

Assignment 5 will be to write a program using Praat's built-in routines to analyze an audio dataset and create a vowel plot, as commonly done in phonetic/sociophonetic analyses.

Class meetings

These will be hands-on, problem-oriented sessions. Students are strongly encouraged to bring a laptop, so that they can investigate and solve problems during in-class activities.

Grade components

- Assignment (90%) will be assigned at the end of each unit (every 2 or 3 weeks) and students will have one week to complete them. There will be five assignments in total.

- Attendance (10%) is included as part of the final grade. "Attending" means coming to class, paying attention, verbally participating and completing in-class exercises.

Grading Scale: Standard OSU grading scheme

A	93–100	A-	90–92		
B+	87–89	B	83–86	B-	80–82
C+	77–79	C	73–76	C-	70–72
D+	67–69	D	60–66		
E	0–59				

Requirements
The class is for 3 credits. Students are expected to attend all classes, participate in all class activities, complete all assigned readings, show evidence of having completed the readings during class discussion, and complete at least four of the five assignments.

Academic Misconduct
It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct.

Students with disabilities that have been certified by the Office for Disability Services will be appropriately accommodated and should inform the instructor as soon as possible of their needs. The Office for Disability Services is located in 150 Pomerene Hall, 1760 Neil Avenue; telephone 292-3307, TDD 292-0901.